

**REMARKS/ARGUMENTS**

Claims 1-15 have been presented for examination. By this response, claims have been amended to remove unneeded limitations and to clarify element in specific contexts to more clearly articulate the invention and to reduce possible ambiguities. Reconsideration is respectfully requested in view of the teachings of the Specification, the limitations in the claims and the teachings of the prior art, as well as the following explanation.

The Examiner's attention is directed to an Information Disclosure Statement filed on March 1, 2004 citing numerous nonpatent references and providing background on the field of the subject invention. The Applicants respectfully request that the references be made of record.

The Applicants have corrected a typographical error in reference number 104 in Page 9, line 9 by replacing the paragraph.

The Examiner has objected to the Specification as failing to adequately teach the Applicants' claimed invention. As reasoning, the Examiner has stated that the Specification lacks the technical details that is "normally associated with such an invention." The Examiner is also evidently of the opinion that the Specification in conjunction with the drawings do not adequately disclose a "detailed technical analysis" of the Applicants' invention.

The Applicants respectfully take strenuous issue with the Examiner's position and request reconsideration. The statutory requirements of the U.S. Patent Laws have been met. It is stated in 35 U.S.C. 112, paragraph 1, that the specification shall contain a written description of the invention ... in such *full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains ... to make and use the same. ...*" The Applicants are confident that the invention stated in the Specification and recited in the claims is straightforward to understand and implement, given the state of the relevant art as understood by the inventors and their peers. The patent Specification is based on the technical disclosure provided by the inventors, some of whom are Ph.D. professionals and professors, to their technician/programmers, providing an outline of the invention, its operation and its basic flow charts to enable one of even less than ordinary skill in this art to make the invention. It is believed that the programmers had no

particular difficulty either in understanding the disclosure or in coding the flow charts presented in accordance with this Specification into commercial products. Under the circumstances, it is respectfully submitted that no further detail is required, such as extensive software listings, more detailed flow charts or explanation outside of the vocabulary necessary to an understanding by those persons of typically ordinary skill.

Nevertheless, the Applicants provide hereinafter a brief narrative to explain the context of invention in simplified terms. It is respectfully submitted that no modification of the specification should be necessary or dictated.

This invention is in the context of an improvement on methods used to extract data from database or documents that often do not share a common origin, structure, or vocabulary. Examples are the entire Internet, various jobs databases, classified newspaper advertising, the inventory of a retailer such as Amazon.com, or an auction service such as eBay. The key concept disclosed in the Specification involves using a "global" classifier that is based on global regularities to "bootstrap" the training of a second or local "local" classifier that uses local regularities that are more effective than the global regularities alone. As an example, a global classifier may be used to extract person-names based for instance on the regularity in patterns for names such as capitalization, common first names, common surnames, and the like. A local classifier may use HTML formatting conventions of résumé pages on a particular web site. The global classifier assists the local classifier to understand whether its data is a person-name

Combining the set of conclusions of the global classifier and of the local classifiers (e.g., through voting, or merely by simply extracting the conclusions of the local classifier once it was trained and ignoring the conclusions of the global classifier in the final classification) on the selected subset of web pages often yields better extraction results than just using a global classifier when local regularities (such as HTML formatting conventions) are used consistently across a domain to identify the item being searched for.

An additional concept disclosed in the Specification and which could admit to further claim development is the concept that a global classifier itself can be re-trained based upon the results produced by many local classifiers. This is a further refinement of the invention

which is believed to be separately novel and patentable, but any use of this teaching is also covered by the present claims.

In view of the foregoing and in view of the level of skill in the art, the Applicants submit that the Specification is adequate to support the claims and that the claims are by no means confusing to one of ordinary skill in the art.

The Examiner has stated that Patent Office personnel are to give claims their "broadest reasonable interpretation." While the Applicants welcome an interpretation that is broad, they are also mindful that the interpretation must be reasonable. The Applicants therefore respectfully suggest that the interpretation evidently applied by the Examiner may well be unreasonable, which has led the Examiner to assert that the claims are confusing. A reasonable interpretation should assuage any concern over confusion.

Claims 1-15 stand rejected under 35 U.S.C. 103(a) as being unpatentable over Chakrabarti in view of Tresch. Chakrabarti has been cited as if it teaches all steps of claim 1 except second classifier training. Tresch has been cited as if it taught training a second classifier using first tentative labels.

The Applicants respectfully submit that Chakrabarti and Tresch are not relevant and have been improperly combined. A review of both Chakrabarti and Tresch reveals that they involve the classification of documents into classes (i.e. whether it's a resume, directory, table-of-contents, product page, 10-Q disclosure document, patent application, etc.). By contrast, the present invention is specific to classification of elements within a webpage, which is a type of document that can form a dataset or record. For specific examples, this invention would classify fields (i.e., the person, place, organizations, salary, etc. fields) within the page using global and local classification. More generically, this invention is more directed to the solution of the problem of extraction by tagging. The present invention involves actually training a separate classifier on separate features for the local classification, and training and running that local classifier over a mere subset of for example documents, e.g., WebPages, that are actually expected to share local regularities.

Chakrabarti definitely lacks any teaching of training a second, namely local classifier from the results of the first, namely global classifier. With this, the Applicants and the

Examiner agree. Where the Applicants dispute the Examiner is the extension of Chakrabarti by reference to Tresch. The Applicants submit that there is nothing in Tresch nor Chakrabarti that suggests such a combination. Moreover, the combination of Chakrabarti and Tresch still does not suggest what may be termed the staging of classifiers or extractors.

Tresch merely teaches iteration, incremental training and quick retraining of the *same* classifiers without any reference to global or local regularities. The combination with Chakrabarti to achieve the claimed invention therefore does not logically follow. In order to make the linkage, Tresch would have to suggest the idea of trying to identify a subset of pages (from the same domain) that is expected to share "local" regularities that are different from the general "global" regularities, and then train at least one and probably many second (local) classifiers over those pages using local regularities as features. The inventive concepts based on the differences between global regularities and local regularities are inherent in the present claims. The claims recite step by step the basic method according to the invention. Furthermore, additional multiple "levels" of regularities (e.g., global, regional, local, sub-local) and the idea of nesting is recited in claim 15. Nothing in Chakrabarti and Tresch establishes that extension. Those claimed concepts are simply lacking in the strained hindsight combination of Chakrabarti and Tresch.

Based upon the Examiner's citation to references in the passages in Tresch and the deficiencies of Chakrabarti, the Applicants question why the Examiner has even applied Tresch. It is therefore respectfully submitted that claim 1 and the other independent claims 8 and 15, clearly define patentable subject matter.

Furthermore, it is submitted that the Tresch reference has been applied out of context to the claimed invention. The rejections of the dependent claims are illustrative.

Claim 2 stands rejected for the same reasons as claim 1 and further based on misplaced assumptions about teachings in Tresch respecting retraining. The Examiner has asserted that it would have been obvious to retrain the second classifier with the second tentative labels on the assumption that the second tentative labels reflect the same labeling.

The Applicants point out that claim 2 does not recite retraining: it recites making a decision regarding retraining. That is a different issue which is not suggested by the passage

cited by the Examiner. This is a decision, not a mere execution of a process to completion. It is submitted that the art of record fails to address the claimed limitation.

Claim 3 stands rejected under the same reasoning applied against claim 2. While claim 3 does relate to training, in claim 3, the type of training is specified: not only is the second classifier trained with the first tentative labels, it is trained using the second tentative labels. Tresch fails to do either or both. It is submitted that the art of record fails to address the claimed limitation.

Claim 4 has been rejected on grounds that it would have been obvious to employ an algorithm that first trains a classifier with training data, and then trains a "new (second)" classifier using the results from the first training set. The claim recites collecting permanent labels associated with elements of a candidate subset of web pages and retraining the first classifier in response to the permanent training labels. This is a specific form of tagging, an extension of the inventive concept that is nowhere taught or suggested in the art, except in the present invention. Nowhere before has it been disclosed that a global classifier be retrained based on the permanent results (e.g., labels) of (many) local classifiers. While this is not a primary contribution of the present invention, it is a contribution that is clearly patentably distinct from even the disclosed and claims subject matter of claim 1.

Claim 5 stands rejected under the same reasoning as the rejection of claim 4. Claim 5 recites that the second (local) classifier treats selected first (global) regularities differently than the first (global) classifier treats the first (global) regularities such that the second (local) regularities *contradict* said first (global) regularities. This is a totally different concept from merely *distinguishing* between "document classes," which is the basis of the citation in the Tresch reference. A contradiction, as explained in the Specification is a special condition. Mutual incompatibility, rather than mere differences, is the basis of this nonobvious advance in the art. Claim 5 stands a patentably distinguishable of the applied art and even over the subject matter of claim 1.

Claim 6 stands rejected using the same reasoning as the rejection of claim 4 and 5. Claim 6 recites a mechanism whereby the contradiction step recited in claim 5 is implemented. The applied reference of Tresch is wholly silent on this point and is thus deficient as a reference.

Claim 7 stands rejected under the same reasoning as applied against of claims 4, 5 and 6. The citation of text in the Tresch reference appears to the Applicants to be out of context and irrelevant to the claimed subject matter. It is a reference to a confidence measure, which is perhaps more reminiscent of claim 8, 11 or 15, where confidence is rated and an output is made with confidence labels. However, it does not apply in any explicit way to the terms of claim 7. The Applicants submit that no claim interpretation of reasonable breadth would encompass the disclosure applied and lead one of ordinary skill in the art to relate that passage to the concept in the output step of combining training results of a global classifier and a local classifier. Confidence labels are not generally thought of as training "results." Claim 7 defines patentable subject matter.

Claim 8 stands rejected much like claim 1 has been rejected. The Examiner has recognized a deficiency in Chakrabarti. Even so, the Applicants submit that this distinguishing step can be as readily recited as a rating step, as amended. (The same amendments were made in claims 11 and 15.)

The Tresch reference, for the reasons previously mentioned, as well as its current application, does not apply. In this context, the cited passage in Tresch appears to be out of context in that changing a classification of a document based on confidence measure is not the same as converting tentative labels to so-called confidence labels which are inherently associated with local regularities. Claim 8 is clearly distinguishable over the art for at least this reason.

Claim 9 stands rejected on grounds that Tresch teaches human-based sorting into classifications. The Examiner appears to have incorrectly equated classifications with regularities. In the present invention, regularities are input, not documents that are typical of a class. It is inferred that the Examiner believes that the human input of documents substitutes for the human input of characteristics of documents, which is the closest evident analogy to regularities in this context. The Applicants point out that the claim specifically recite the inputting of descriptions of global regularities, which is a specific type of input. It is respectfully submitted that claim 9, without more, patentably defines invention over the art of record.

Regarding claim 10, Chakrabarti has been cited for teachings related to application of inductive operations, pointing to Col. 12, lines 36-40. Therein Chakrabarti recites

the use of class models and Bayes law applied to document classification. Again, there is nothing used therein that applies to categorizing data within a document so the data can be compared with data in other documents where there is no shared origin, structure or vocabulary. Claim 10 stands as patentably distinct.

Claims 11-14 stand rejected under the same rationale as the rejection of claim 10. The Applicants submit that the rationale is deficient for the reasons previously stated.

Claim 15 stands rejected under the same rationale as the rejections of claims 1 and 8. The Applicants submit that the rationale is deficient for the reasons previously stated. In addition, the Applicants point out that claim 15 recites additional steps of a multiple level of hierarchies between global and local regularities and a recursive process of training whereby the more global regularities are trained by the more local regularities. This is nowhere taught or suggested in the applied art. Moreover, it is a nonobvious extension that is distinguishable over the processes recited in the preceding claims. For these reasons, the Applicants submit that this claim defines separately patentable subject matter.

### CONCLUSION

In view of the foregoing, Applicants believe all claims now pending in this Application are in condition for allowance. The issuance of a formal Notice of Allowance at an early date is respectfully requested.

If the Examiner believes a telephone conference would expedite prosecution of this application, please telephone the undersigned at 650-326-2400.

Respectfully submitted,



Kenneth R. Allen  
Reg. No. 27,301

TOWNSEND and TOWNSEND and CREW LLP  
Two Embarcadero Center, Eighth Floor  
San Francisco, California 94111-3834  
Tel: (650) 326-2400  
Fax: (650) 326-2422  
KRA:dim

60100761 v1